Abstract Title Page
Not included in page count.

Title:
When Does Teacher Incentive Pay Raise Student Achievement? Evidence from Minnesota's Q-Comp Program

Author(s):
Aaron Sojourner, U Minnesota, Department of Human Resources and Industrial Relations, asojourn@umn.edu

Kristine West, U Minnesota, Department of Applied Economics

Elton Mykerezi, U Minnesota, Department of Applied Economics

Abstract Body

Limit 4 pages single spaced.

Background / Context:

Teachers vary widely in their ability to produce student achievement gains (Hanushek 1971, Hanushek and Rivkin 2010) but this ability is not predicted by educational degrees or experience beyond the first few years of a teacher's career (Hanushek 2003, Aaronson et al 2007). This has large economic consequences (Chetty et al 2010, Hanushek 2010), which motivates policy and research interest in pay for performance (P4P). Advocates of P4P believe that tying teacher compensation to performance will support increased efforts from incumbent teachers and attract better potential teachers to the profession (Lazear 2003). Many school districts and states are experimenting with P4P plans, which set compensation criteria beyond the conventional ones: experience and education. Plans differ on many dimensions including whether teachers are rewarded individually or in teams, whether for test scores or other measures of teacher quality, and in the magnitude of incentive pay available.

Empirical evidence on the relative and absolute merit of these programs is decidedly mixed. While reviews of the literature point to some gains from P4P (Springer and Podgursky 2007, Neal 2011), evaluations of two large-scale P4P plans that were implemented as randomized trials found null or even negative effects on student achievement (Springer et al 2010, Fryer 2011). Whether plans implemented as long-term policies rather than short-term experiments or plans with other designs would produce better results remains an open question of great interest.

Purpose / Objective / Research Question / Focus of Study:

We provide new evidence on several issues of theoretical importance related to P4P contracts in education. For instance, it is not clear what the optimal team size for targeting bonuses should be. On one hand, incentives tied to school-level criteria may encourage efficient effort if there are positive externalities from cooperation (Weitzman and Kruse 1990) or variations in incentive strength across teachers (Ahn 2008). On the other hand, free riding may make individual or small team incentives preferable (Kandel and Lazear 1992). Since Q-Comp districts adopted a wide range of P4P contracts, we are able to investigate whether incentives offered at lower levels of aggregation (such as the individual teacher or grade) are more or less productive than those offered at higher levels of aggregation (such as the school or district level). There are also important theoretical questions about how to measure teacher quality and performance. Measures based on principal or peer subjective evaluations have received some attention in the literature, especially since principals seem able to identify the best and worst teachers (Jacob and Lefgren 2008). However, high-stakes subjective evaluation processes may be captured and converted into de facto salary augmentations (Neal 2011). Minnesota's Q-comp offers a valuable opportunity to examine if a high-stakes P4P plan based on subjective evaluations affects educational outcomes.

Setting:

The State of Minnesota implemented its Quality Compensation (Q-Comp) program in 2005 as the signature education initiative of Governor Tim Pawlenty. The Minnesota Department of Education (MDE) set general guidelines for acceptable programs and invited districts to propose specific P4P program designs that they would implement. If the proposal was approved, the state authorized additional funding to the district. With its Race to the Top Fund and Teacher Incentive Fund, the U.S. Dept. of Education has adopted a similar policy for disbursing billions.

Districts designed plans that varied along many dimensions. Each district was required to specify the maximum incentive pay they would make available to teachers based on different types of criteria and there is great variation in what they chose. This allows us to construct continuous measures of each district's plan in terms of dollars at stake for: (1) individual teacher actions or outcomes, (2) school wide goals, or (3) through a subjective evaluation process. We exploit this variation in P4P plan designs, along with variation in when districts adopted Q-Comp, to provide evidence on the effect of plan design features on achievement scores and other outcomes.

For a number of reasons, Q-Comp provides an excellent opportunity to learn about the effects of P4P in a policy framework mirroring recent national efforts. First, the program has been in effect for over five years and was implemented as a permanent program rather than a time-limited experiment. Second, there is substantial variation in what criteria trigger P4P bonuses. Third, Minnesota has one of the longest lasting and most widely used inter-district open enrollment policies and a large number of charter schools, so parents have substantial choice in public schooling. This makes it possible to examine the effect P4P designs have on net student movements, which can reflect changes in parent demand. Overall, understanding the Minnesota P4P experience can provide valuable information to policy makers nationwide.

Population / Participants / Subjects:
Table 1 describes the number of districts, schools and students participating and not participating in Q-Comp each year. The population is all Minnesota public schools, including charters each constituting its own "district." In Q-Comp's first year, 2005, only eight of the state's 504 districts participated (1.6%). These included 33,674 of the 838,997 students (4.0%). New cohorts adopted each year. By the 2009, 14.1% of districts with 28.6% of students participated. Most analysis will focus on grades 3 to 8 because, in these grades, all students took both math and reading tests. Participation statistics are provided for schools in this sample in the bottom panel
.

Intervention / Program / Practice:
Q-Comp is a package of reforms with P4P at its center. We focus our attention on the performance pay component because these are the most interesting theoretically, the best measured in the available data, and the most likely to constitute a real policy change from the pre-adoption period. Data on performance pay are collected primarily from letters sent by the MDE to each district upon approval of its Q-Comp application. The letters detail agreed-upon features of the plan. From these, we create three variables for each district measuring the maximum performance pay available to teachers for the following types of criteria:

- **Teacher P4P$**: anything under a teacher's primary influence. This includes inputs related to professional development (e.g., attending meetings, taking classes, completing professional development plans and self-assessments) and outputs close to the teacher (e.g., student performance on teacher-created assessments, a teacher's own students' standardized test scores). It also includes analogous small team or grade-wide outcomes. The descriptions in the approval letters do not allow us to consistently distinguish between various elements within this domain.
- **School P4P$**: anything linked to school-wide or district-wide outcomes. These primarily involve hitting standardized test score targets.

- **Evaluation P4P$**: anything linked to classroom observations or subjective evaluations performed by peers, administrators, or a district-sanctioned mentors including a formal annual review process.

Districts vary in the total levels of pay available across these three dimensions as well as the shares available through each dimension. The value of these variables is shared by all school-grades within a district-year. Figure 1 illustrates and Table 2 summarizes the distribution of these measures across participating districts.

Research Design:

To learn about the impact of Q-Comp on student achievement, we analyze a panel of student achievement, demographic, and school characteristic data defined at the year-district-grade level using generalized difference-in-difference methods. We study how districts' student achievement changes as their Q-Comp participation changes and depending on the design of the P4P program they adopt. For each academic year indexed t=2005, 2006,...2009, in each school district indexed d=1,2...D, in each tested grade indexed g=3,4,...8, and in either math or reading indexed by b, we use variants of this model:

$$y_{tdgb} = \beta_{gb} Q_{tdgb} + \rho y_{(t-1)d(g-1)} + \alpha_{gb} w_{tdg} + \gamma_{dgb} + \delta_{tgb} + \varepsilon_{tdgb}$$

Interest centers on the effects of Q-Comp participation and of features of the P4P designs adopted. These are captured by β_{bg}. Specification (A) treats the whole pre-adoption period as the reference category. Specifications (B) conditions on and measures pre-adoption differences in achievement levels between adopters and non-adopters using a dummy to indicate observations that come from years more than one year pre-adoption. To measure the effects of various aspects of Q-Comp P4P program design, we use different definitions of *Q*. We define *Q* to be a vector measuring P4P design features in years that they are participating and zeros otherwise. In particular, we measure maximum P4P bonus available to teachers in a district for three kinds of criteria, measured as Teacher P4P$, School P4P$, and Evaluation P4P$.

Since Q-Comp participation is not randomly assigned, there may be systematic unobserved differences between districts that influence both Q-Comp adoption and our outcomes, which would bias estimates of program effects. We use four main strategies to guard against this threat.

First, since average student achievement may vary over time due to differences in student cohorts within any given district and grade, we condition on lagged math and reading achievement ($y_{(t-1)d(g-1)}$) and student demographic characteristics and school-level variables (w_{tsg}).

Second, district-grade fixed effects are included to remove time-invariant, additive unobserved differences in achievement levels γ_{dgb} between schools. The model is identified from within-district-grade-subject, across-time variation. Fixed effects for each year-grade-subject are also included. These terms identify counter-factual year effects for each grade and subject δ_{tgb}. The comparison group matters here because their experience across years defines these time effects. This is a generalization of difference-in-difference analysis that relies on differences in the timing of adoption across districts to separate time effects from program effects. Within the restrictions of functional form, this model yields unbiased estimates of program effects if selection into Q-Comp is based on stable differences in achievement levels. The crucial assumption is that within-district, time-varying, unobserved influences on achievement levels are not systematically related to whether or when a district adopted Q-Comp or the features of the design it adopted. The estimates of β may be biased if districts select into participation or design based on fluctuations in achievement levels.

Third, we estimate these models with three different comparison groups. We compare the experience of participants to that of either (1) all other districts in the state, (2) districts that applied to Q-Comp but failed to adopt, due either to the state rejecting the proposal or their teachers voting against it, and (3) just Q-Comp adopters who have not yet adopted.

Fourth, since the model is identified by assuming exogenous timing of adoptions, we drop each cohort and assess whether the results change. They generally do not.

Data Collection and Analysis:
Data were collected primarily from the Minnesota Department of Education and aggregated to the district-year level. Our primary achievement measure is the Minnesota Comprehensive Assessments Series II (MCA-II) average scores, standardized to mean 0 and standard deviation 1 across schools within year-grade-subject. Demographics, school and district characteristics were also available. Q-Comp designs for each participating district were coded from the state's program approval letters to each district.

Findings / Results:
Districts that tied P4P$ to teacher level actions or outcomes produced large effects on reading and math growth, as described in Tables 8 and 10. For every $1000 in Teacher P4P$ offered to teachers, districts experienced an additional 0.172 (0.068) σ increase in reading growth and an additional 0.136 (0.061) σ increase in math growth. These are large and cheap effects. Districts that tied P4P$ to school wide achievement outcomes or subjective evaluations experience null or negative changes in growth. These are robust to cohort exclusions, as in Table 16.

These countervailing effects cancel each other out in the aggregate. The average effect of Q-Comp participation is null in math and in reading. In our preliminary results in Table 11, we do not observe much change in student demand or teacher sorting.

Conclusions:
The experience in Minnesota adds to our understanding of locally-designed P4P plans. The grantor-grantee relationship between education authorities and districts has advantages because it allows use of local information and experimentation in finding appropriate, feasible designs. Our findings suggest that if a granting authority proposes a range of reforms and allows districts to design plans locally, many districts (in cooperation with local teachers' unions) will design plans that base rewards largely on subjective evaluations and this does not seem to benefit student achievement. On the other hand, some districts (in cooperation with their local teachers' unions) will weight rewards to more specific teacher-centered criteria and this appears beneficial for achievement.

The fact that, despite large gains in some areas of the program, Minnesota spent $200 million to get a net effect of zero also points out risks associated with too much local control over the plans.

Some plans will operate to extract rents from the state more than to improve education. State and federal governments can, however, use the experiences of early adopters, such as Minnesota, to chose more appropriate program guidelines.

Appendix A. References

1. Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. Teachers and Student Achievement in Chicago Public High Schools. *Journal of Labor Economics*, 25(1).

2. Ahn, Tom. 2008. The Missing Link: Estimating the Impact of Incentives on Effort and Effort on Production Using Teacher Accountability Legislation. Duke University Working Paper.

3. Atkinson, Adele, Burgess, Simon, Croxson, Bronwyn, Gregg, Paul, Propper, Carol, Slater, Helen and Deborah Wilson. 2009. Evaluating the Impact of Performance-related Pay for Teachers in England. *Labour Economics*, 16(3): 251–61.

4. Baker, George. 1992. Incentive Contracts and Performance Measurement. *Journal of Political Economy*, 100: 598–614.

5. Baker, George, Robert Gibbons, and Kevin J. Murphy. 1994. Subjective Performance Measures in Optimal Incentive Contracts. *Quarterly Journal of Economics*, 109, 1125–56.

6. Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2010. How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR. National Bureau of Economic Research Working Paper No. 16381.

7. Fryer, Roland. 2011. Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. NBER Working Paper 16850.

8. Glazerman, Steven and Allison Siefullah. 2010. An Evaluation of the Teacher Advancement Program (TAP) in Chicago: Year Two Impact Report. Mathematica Policy Research, Inc.

9. Hanushek, Eric. 1971. Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data. *The American Economic Review*. 61(2): 280-288.

10. Hanushek, Eric. 2003. The Failure of Input-Based Resource Policies. *The Economic Journal.* 113(485): F64-F98.

11. Hanushek, Eric. 2010. The Economic Value of Higher Teacher Quality. National Bureau of Economic Research Working Paper 16606.

12. Hanushek, Eric. and Steven Rivkin. 2010. Generalizations about Using Value-Added Measures of Teacher Quality. *American Economic Review.* 100(May 2010):267-271.

13. Hezel Associates, LLC. 2009. Quality Compensation for Teachers Summative Evaluation.

14. Holmstrom, Bengt, and Paul Milgrom. 1991. Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, and Organization*, 7: 24-52.

15. Jacob, Brian and Lars Lefgren. 2008. Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. *Journal of Labor Economics*, 26(1):101-36.

16. Jacob, Brian. 2010. Do Principals Fire the Worst Teachers? National Bureau of Economic Research Working Paper 15715.

17. Johns, Emily. 2009. Is It "Merit Pay" If Nearly All Teachers Get It? *Minneapolis Star Tribune.*

18. Kandel, Eugene and Edward Lazear. 1992. Peer Pressure and Partnerships. *Journal of Political Economy*, 100(4): 801–17.

19. Lavy, Victor. 2002. Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement. *Journal of Political Economy*, 110(6): 1286–1317.

20. Lazear, Edward P. 2003. Teacher Incentives. *Swedish Economic Policy Review*, 10.

21. Lovenheim, Michael. 2009. The Effect of Teachers' Unions on Education Production: Evidence from Union Election Certifications in Three Midwestern States. *Journal of Labor Economics*, 27(4):525-87.

22. Martins, Pedro. 2009. Individual Teacher Incentives, Student Achievement and Grade Inflation. IZA Discussion Paper No. 4051.

23. Muralidharan, Karthik and Venkatesh Sundararaman. 2011. Teacher Performance Pay: Experimental Evidence from India. *Journal of Political Economy*. 119(1).

24. Nadler, Carl and Matthew Wiswall. 2011. Risk Aversion an Support for Merit Pay: Theory and Evidence from Minnesota's Q Comp Program. *Education Finance and Policy*, 6(1):75-104.

25. Neal, Derek. 2011. The Design of Performance Pay in Education. NBER Working Paper No. 16710.

26. Nobels, James. 2009. Evaluation Report: Q Comp Quality Compensation. Minnesota Office of the Legislative Auditor.

27. Podgursky, Michael and Matthew Springer. 2007. Teacher Performance Pay: A Review. *Journal of Policy Analysis and Management*, 26(4):909-50.

28. Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger. 2008. Can You Recognize An Effective Teacher When You Recruit One? NBER Working Paper 14485.

29. Springer, M., D. Ballou and Peng. 2008. Impact of the Teacher Advancement Program on Student Test Score Gains: Findings for an Independent Appraisal. National Center on Performance Incentives Working Paper 2008-19.

30. Springer, Matthew, Laura Hamilton, Daniel McCaffrey, Dale Ballou, Vi-Nhuan Le, Matthew Pepper, J.R. Lockwood and Brian Stecher. 2010. Teacher Pay for Perfor-

mance: Experimental Evidence from the Project on Incentives in Teaching. National Center on Performance Incentives.

31. Taylor, Eric and John H. Tyler. 2011. The Effect of Evaluation on Performance: Evidence from Longitudinal Student Achievement Data of Mid-Career Teachers. National Bureau of Economic Research Working paper 16877.

32. Taylor, Lori and Matthew Springer. 2008. Optimal Incentives for Public Sector Workers: The Case of Teacher-Designed Incentive Pay in Texas. National Center on Performance Incentives Working Paper 2009-05.

33. Tyler, John, Eric Taylor, Thomas Kane, and Amy Wooten. 2010. Using Student Performance Data to Identify Effective Classroom Practices. *American Economic Review*, 100(2): 256-60.

34. Vigdor, Jacob L. 2008. Teacher Salary Bonuses in North Carolina. Conference paper, National Center on Performance Incentives.

35. Weitzman, Martin L. and Douglas Kruse. 1990. Profit sharing and productivity. In *Paying for Productivity*. Edited by A. Blinder. Brookings Institution: Washington, D.C.

Appendix B. Tables and Figures

Not included in page count.

Table 1: District and School Q-Comp Participation by Year

	Participants			Non-Participants		
Year	Districts	Schools	Students	Districts	Schools	Students
			All schools			
2005-06	8	59	33,674	496	2,197	805,323
2006-07	50	322	183,216	458	1,922	657,346
2007-08	60	397	231,465	456	1,856	606,113
2008-09	70	429	252,716	457	1,786	583,218
2009-10	74	411	239,489	451	1,796	597,141
			Schools including at least one grade in 3 to 8			
2005-06	7	52	23,131	404	1,511	567,202
2006-07	36	255	129,754	379	1,338	463,862
2007-08	43	309	162,499	379	1,278	462,980
2008-09	52	328	176,870	381	1,258	413,023
2009-10	56	315	166,697	375	1,256	427,549

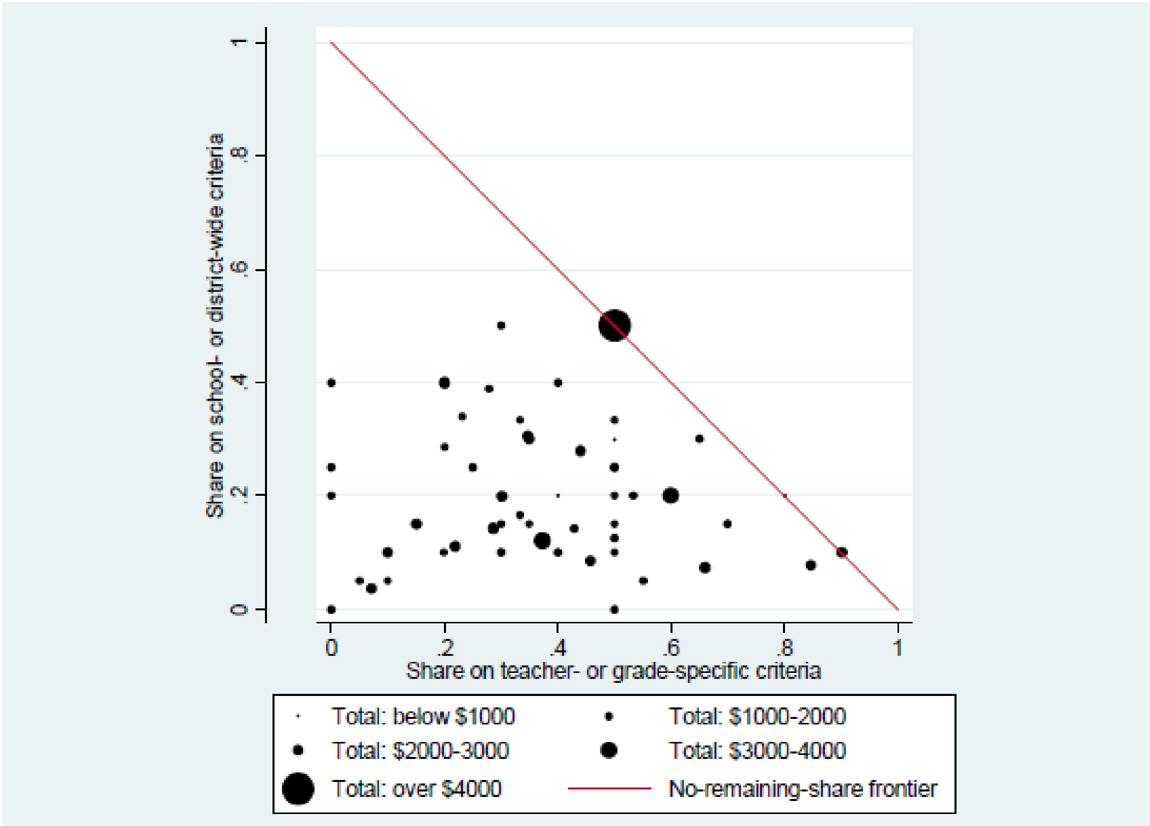

Figure 1: Joint distribution of P4P designs across Q-Comp districts

Table 2: Summary statistics for district Q-Comp program design variables measuring maximum pay available through each dimension, in thousands of dollars

	Unweighted		Weighted by students			
	Mean	Std. Dev.	Mean	Std. Dev.	Min.	Max.
Teacher P4P$	0.682	0.585	0.815	0.652	0	2.5
School P4P$	0.334	0.342	0.247	0.225	0	2.5
Evaluation P4P$	0.727	0.58	0.987	0.743	0	2.5
Number of participating districts			77			

Note: the 2010 cohort included additional districts but their plans are not coded.

Table 8: Program design effects on student achievement *growth* - reading

DV: Reading average achievement for district-grade-year						
Sample	Full		Interested Only		Adopters Only	
Specification	(A)	(B)	(A)	(B)	(A)	(B)
Teacher P4P$	0.17**	0.172**	0.172**	0.173**	0.206***	0.208***
	(0.068)	(0.068)	(0.069)	(0.068)	(0.068)	(0.068)
School P4P$	-.297	-.314*	-.351*	-.353*	-.352*	-.361*
	(0.191)	(0.19)	(0.192)	(0.192)	(0.191)	(0.19)
Evaluation P4P$	-.050	-.051	-.052	-.052	-.042	-.043
	(0.046)	(0.046)	(0.051)	(0.051)	(0.052)	(0.052)
1(Missing P4P$)	-.161***	-.163***	-.176***	-.176***	-.125*	-.125*
	(0.049)	(0.049)	(0.061)	(0.061)	(0.069)	(0.068)
Lagged reading	0.351***	0.351***	0.328***	0.328***	0.312***	0.312***
	(0.017)	(0.017)	(0.034)	(0.034)	(0.038)	(0.038)
Lagged math	0.133***	0.133***	0.161***	0.161***	0.163***	0.163***
	(0.016)	(0.016)	(0.031)	(0.031)	(0.037)	(0.037)
2+ pre-adoption		-.022		-.002		-.014
		(0.044)		(0.042)		(0.045)
Student observables	Yes	Yes	Yes	Yes	Yes	Yes
District-grade FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-grade FE	Yes	Yes	Yes	Yes	Yes	Yes
N districts	446	446	132	132	98	98
N students	1339042	1339042	578414	578414	446951	446951
Adjusted R^2	0.934	0.934	0.964	0.964	0.953	0.953

Coefficient (within-district SE). Significance: $*$: 10% $**$: 5% $***$: 1%.

Variables are year-district-grade averages.

Lags are prior year, prior grade $(t-1)d(g-1)b$.

Table 10: Program design effects on student achievement *growth* - math

DV: Math average achievement for district-grade-year						
Sample	Full		Interested Only		Adopters Only	
Specification	(A)	(B)	(A)	(B)	(A)	(B)
Teacher P4P$	0.132**	0.136**	0.113*	0.121*	0.112*	0.123*
	(0.061)	(0.061)	(0.064)	(0.063)	(0.065)	(0.065)
School P4P$	-.275	-.312*	-.245	-.288	-.248	-.287
	(0.175)	(0.177)	(0.183)	(0.182)	(0.183)	(0.181)
Evaluation P4P$	-.046**	-.049**	-.042	-.044*	-.044	-.044
	(0.024)	(0.023)	(0.027)	(0.026)	(0.028)	(0.027)
1(P4P$ missing)	0.043	0.037	0.028	0.023	0.043	0.041
	(0.067)	(0.064)	(0.081)	(0.077)	(0.093)	(0.089)
Lagged reading	0.181***	0.181***	0.168***	0.167***	0.153***	0.152***
	(0.018)	(0.018)	(0.035)	(0.035)	(0.042)	(0.042)
Lagged math	0.341***	0.341***	0.376***	0.377***	0.397***	0.398***
	(0.018)	(0.018)	(0.035)	(0.035)	(0.04)	(0.04)
2+ pre-adoption		-.050		-.068*		-.068*
		(0.036)		(0.037)		(0.039)
Student observables	Yes	Yes	Yes	Yes	Yes	Yes
District-grade FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-grade FE	Yes	Yes	Yes	Yes	Yes	Yes
N districts	414	414	127	127	93	93
N students	1038147	1038147	450596	450596	347832	347832
Adjusted R^2	0.936	0.936	0.963	0.963	0.957	0.957

Coefficient (within-district SE). Significance: $*$: 10% $**$: 5% $***$: 1%.

Variables are year-district-grade averages.

Lags are prior year, prior grade $(t-1)d(g-1)b$.

Table 11: Program design effects on alternative outcomes using academic years from 2003-2004 to 2009-2010 and all available grades (K-12)

	Teacher			Student		
	Log (mean pay)	Mean yrs. exper.	% M.A.	Log (Enrlmt.)	Inter-dist. net flow %	Attend. %
Teacher P4P$	0.025**	-.266	1.621*	0.022	-.392	-.062
	(0.012)	(0.211)	(0.887)	(0.04)	(0.766)	(0.074)
School P4P$	-.097	0.211	-2.806	0.139	3.301*	-.082
	(0.069)	(1.005)	(3.529)	(0.113)	(1.793)	(0.375)
Evaluation P4P$	0.025***	0.139	0.393	-.059**	0.418	-.058
	(0.01)	(0.151)	(0.383)	(0.024)	(0.33)	(0.089)
2+ pre-adoption	0.00007	0.279	-.141	-.087***	-1.136*	0.05
	(0.007)	(0.224)	(0.441)	(0.027)	(0.587)	(0.152)
Excludes:					Charters	'06,'09
N district-years	498	500	500	558	345	543
Weighted by	FTE	FTE	FTE	Grade	Student	Student
N	356992	357307	357307	3974	5749030	4132170
Adj R^2	0.921	0.887	0.948	0.986	0.907	0.884

Significance: $*$: 10% $**$: 5% $***$: 1%.

Coefficient (within-district SE). Year effects and district effects included. All use district-level variables, except enrollment (district-grade) & attendance rate (school-grade).

Table 16: Robustness of growth model to dropping any adoption cohort - pooled across grades 3 to 8 and academic years 2005-06 to 2009-10

	Adoption cohort excluded from analysis:					
	2005	2006	2007	2008	2009	2010
			Reading			
Teacher P4P$	0.166**	0.284***	0.151**	0.168**	0.177**	0.172**
	(0.068)	(0.073)	(0.076)	(0.077)	(0.069)	(0.067)
School P4P$	-.305	-.479***	-.266	-.338	-.327	-.322*
	(0.19)	(0.181)	(0.193)	(0.277)	(0.201)	(0.19)
Evaluation P4P$	-.055	-.012	-.058	-.056	-.052	-.054
	(0.047)	(0.025)	(0.05)	(0.06)	(0.047)	(0.046)
1(Missing P4P$)	-.170***	-.054*	-.190***	-.163***	-.161***	-.168***
	(0.05)	(0.029)	(0.066)	(0.048)	(0.049)	(0.049)
2+ pre-adoption	-.021	0.008	0.002	-.060	-.018	-.036
	(0.044)	(0.044)	(0.045)	(0.047)	(0.045)	(0.062)
N districts	439	407	438	436	440	419
N district grades	2829	2565	2821	2832	2874	2750
N tested students	1292480	1094541	1257031	1301639	1335331	1306279
Adj. R^2	0.934	0.928	0.931	0.931	0.934	0.935
			Math			
Teacher P4P$	0.135**	0.147**	0.132*	0.138**	0.136**	0.136**
	(0.061)	(0.06)	(0.068)	(0.069)	(0.062)	(0.061)
School P4P$	-.309*	-.328	-.294	-.311	-.313*	-.328*
	(0.177)	(0.214)	(0.179)	(0.226)	(0.188)	(0.179)
Evaluation P4P$	-.049**	-.040	-.049*	-.054**	-.047**	-.052**
	(0.023)	(0.042)	(0.026)	(0.022)	(0.023)	(0.023)
1(Missing P4P$)	0.007	0.25***	-.029	0.043	0.038	0.034
	(0.059)	(0.071)	(0.079)	(0.063)	(0.064)	(0.063)
2+ pre-adoption	-.042	-.036	-.032	-.061	-.043	-.072
	(0.036)	(0.039)	(0.041)	(0.045)	(0.037)	(0.046)
N districts	438	406	437	435	439	418
N district grades	2826	2562	2818	2829	2871	2747
N tested students	1249991	1058062	1215480	1258601	1291574	1263516
Adj. R^2	0.924	0.914	0.921	0.921	0.924	0.924

Coefficient (within-district SE). Significance: $*$: 10% $**$: 5% $***$: 1%.

Reading (math) analogous to column 2 of Table 8 (10), except for exclusion of adoption cohorts.